



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

공학석사학위논문

추가 정보를 고려한

상품 리뷰 요약 기법

A Product Review Summarization

Considering Additional Information

2019 년 8 월

서울대학교 대학원

컴퓨터공학부

윤 제 연

초 록

문서 요약이란 주어진 문서로부터 특정 사용자나 작업에 적합하게 축약한 형태의 문서를 생성하는 과제를 의미한다. 인터넷이 발달함에 텍스트를 포함한 여러 종류의 데이터들이 폭발적으로 증가하고 있고, 이러한 상황에서 문서 요약 기술이 지니는 가치는 점점 증대되고 있다. 또한 기계가 언어를 어떻게 이해하는지에 대한 연구의 한 갈래로써 중요한 가치를 지니고 있다.

과거에는 그래프 기반 방법론이나 통계 기반 방법론을 이용한 문서 요약 기법들이 일반적이었으나, 최신의 다양한 딥러닝 기반의 방법론을 이용한 문서 요약 모델들이 주로 연구되고 있다. 이러한 다양한 모델들이 사람이 직접 생성한 요약에 비견될 정도의 좋은 요약 성능들을 보이지만, 학습 데이터들의 양과 질에 따라 성능이 좌우된다는 문제점이 있다.

문서 요약의 대상이 되는 텍스트로는 일반적으로 뉴스 기사, 웹 사이트, 상품 리뷰 등이 있다. 대다수의 연구는 학습 데이터들의 양과 질 모두 좋은 뉴스 기사를 토대로 연구하나, 본 연구에서는 상대적으로 학습 데이터의 양과 질이 열악한 상품 리뷰에 대한 요약 기법을 다루고자 한다.

따라서 본 연구에서는 이러한 데이터의 양과 질의 열악함을 해결하기 위해, 주어진 상품 리뷰들에 대한 추가적인 정보를 더해서 더 좋은 품질의 요약을 제공하는 요약 모델을 제안한다. 본 논문에서 제안한 모델과 기존 모델이 생성한 요약 결과를 비교하고, 또 사용자 평가를 포함한 다양한 실험 결과, 본 연구에서 제안한 상품 리뷰 요약의 품질이 다른 모델보다 관련성과 가독성 모두 더 좋은 품질을 보였으며 추가적인 실험을 통해 추가 정보가 문서 요약에 어떤 영향을 주는 지에 대해 확인했다.

주요어: 문서 요약, 시퀀스-투-시퀀스 모델, 주목 매커니즘, 기억 신경망, 인공 신경망

학 번: 2016-26565

목 차

| | |
|--------------------------------|-----|
| 초 록 | i |
| 목 차 | iii |
| 표 목차 | v |
| 그림 목차 | vi |
| 제 1장 서론 | 1 |
| 1.1 연구의 배경 및 내용 | 1 |
| 1.2 논문의 구성 | 2 |
| 제 2장 관련 연구 | 3 |
| 2.1 시퀀스-투-시퀀스 주목 모델 | 3 |
| 2.2 기억 신경망 | 5 |
| 제 3장 배경 지식 | 7 |
| 3.1 문서 요약 신경망 | 7 |
| 제 4장 추가 정보를 고려한 문서 요약 모델 | 9 |

| | |
|------------------------------|-----------|
| 4.1 기존 연구의 한계점..... | 9 |
| 4.2 추가 정보 인코더 | 10 |
| 4.3 기존 모델과 결합..... | 12 |
| 제 5장 실험 방법 및 결과 | 15 |
| 5.1 데이터 셋 | 15 |
| 5.1.1 학습 데이터 셋..... | 15 |
| 5.1.2 데이터 전처리 | 18 |
| 5.2 실험 설계..... | 18 |
| 5.3 실험 결과 및 분석 | 20 |
| 5.4 추가 정보 영향 평가 | 23 |
| 5.4.1 추가 정보 가중치 실험..... | 23 |
| 5.4.2 추가 정보 제어 실험..... | 25 |
| 5.4.3 추가 정보 영향력 실험..... | 27 |
| 제 6장 결론 및 향후 연구 | 29 |
| 6.1 결론..... | 29 |
| 6.2 향후 연구 | 30 |

| | |
|-------------|-----|
| 참고 문헌 | 3 1 |
|-------------|-----|

| | |
|----------------|-----|
| Abstract | 3 3 |
|----------------|-----|

표 목차

| | |
|--------------------------------|-----|
| 표 1 수집한 데이터 셋에 대한 정보 | 1 5 |
| 표 2 상품 후기 데이터 셋의 추가 정보 표 | 1 6 |
| 표 3 상품 후기 데이터 항목 예시 | 1 6 |
| 표 4 데이터 전처리 예시 | 1 8 |
| 표 5 각 모델의 블루 스코어 | 2 1 |
| 표 6 사용자 평가 결과 | 2 2 |
| 표 7 각 모델별 요약 결과 | 2 2 |
| 표 8 추가 정보 항목별 가중치 평균 | 2 3 |
| 표 8 추가 정보 제어 요약 생성 결과 | 2 6 |

그림 목차

| | |
|----------------------------------|-----|
| 그림 1 키-벨류 기어 신경망의 구조 | 5 |
| 그림 2 추가 정보 인코더의 구조 | 1 0 |
| 그림 3 요약 신경망의 기본적인 구조 | 1 2 |
| 그림 4 추가 정보 임베딩 생성 과정 | 1 3 |
| 그림 5 추가 정보 임베딩이 추가된 디코더 구조 | 1 4 |

제 1장 서론

1.1 연구의 배경 및 내용

문서 요약(Text Summarization)은 긴 텍스트로 이루어진 문서의 정보를 목표 사용자와 목표 작업에 적합하게 압축한 문서를 생성하는 것을 목표로 하는 과제이다. 대량의 정보를 압축해 요약함으로써, 뉴스 요약문 작성, 문서 제목 생성 등 다양한 분야에 사용되고 있다. 또한 문서 요약 기법은 기계 번역(Machine Translation)[2][3][4], 텍스트 질의 응답(Text Question Answering)과 함께 기계 학습(Machine Learning)의 자연어 이해(Reading Comprehension)의 토대가 되는 중요한 분야이다.

문서 요약은 요약 방식에 따라 추출 요약(Extractive Summarization)[8][9]과 생성 요약(Generative Summarization)[10][11]으로 구분한다. 추출 요약은 대상 문서에 이미 존재하는 문장이나 구절 중 의미 있는 부분을 선택하고 추출해서 요약을 하는 방법론이고, 생성 요약은 문서의 내용을 압축해서 새로운 문장들을 생성하는 방법론이다. 과거에는 이 두 가지 방법론 각각에 대해 연구가 이루어졌지만, 현재는 두 가지 방법론을 같이 사용하는 혼합 모델(Hybrid Model)[12][13]에 대한 연구가 주로 이루어지고 있다.

딥러닝 이전의 문서 요약 기법들은 단어 빈도수를 위한 통계 기반[14] 혹은 그래프 기반[10] 기법들의 연구가 주가 되었지만 활용할 수 있는 분야가 적었고, 생성한 요약의 결과가 사람에 의한 요약보다 질이 좋지 않다는 문제가 있었다. 하지만 최근의 시퀀스-투-시퀀스 모델(Sequence-To Sequence-Model)[2], 주목 메커니즘(Attention Mechanism)[4]과 같은 딥러닝 기반의 모델이 등장함으로써 문서 요약 기법의 성능은 사람에 근접한 요약 성능을 보여주고 있다.

하지만 이러한 딥러닝 기반의 모델들은 많은 양의 정제된 학습 데이터가 필요하고, 모델 성능이 데이터의 양과 질에 좌우되는 특징이 있다. 대다수의 연구가 CNN/Daily Mail[1]을 학습 데이터로 사용하고 학습 데이터의 양과 질이 풍족한 상황을 가정하고 있다. 이 때문에 데이터가 풍부한 주제에 대해서는 좋은 결과를 이끌어 내지만 그렇지 못한 학습 데이터들에 대해서는 좋은 성능을 보장하지 못하는 문제점이 있다. 또한 실제 데이터들은 대부분 양과 질이 모두 부족한 것이 현실이다.

따라서 본 연구에서는 양과 질 모두 부족한 상품 리뷰 데이터에서 좋은 요약을 생성하기 위해 추가적인 정보를 이용하는 새로운 상품 리뷰 요약 모델을 제안한다. 기존 모델에 비해서 본 연구에서 제안하는 추가 정보를 이용한 문서 요약 모델이 상품 리뷰 데이터셋에 대해서 더 좋은 요약을 생성함을 보였으며 추가 정보가 요약 생성에 어떤 영향을 주는지 실험을 통해 확인했다.

1.2 논문의 구성

본 논문의 구성은 다음과 같이 이루어져 있다. 2 장에서는 자연 언어 처리 분야의 대표적인 생성 모델과 외부 정보를 이용하는 모델들에 대해 소개하고, 3 장에서는 본 논문에서 기준이 되는 요약 신경망 모델에 대해 설명한다. 4 장에서는 기존 연구의 한계를 보완하는 추가 정보를 활용한 문서 요약 시스템을 제안한다. 5 장에서는 실험을 하기 위해 온라인 쇼핑몰에서 수집한 상품 리뷰 데이터 셋에 대해 소개하고, 다양한 환경에서의 모델들의 성능을 비교 및 평가한다. 마지막 6 장에서는 본 논문의 결론을 맺으며 향후 연구 방향에 대해 논의한다.

제 2장 관련 연구

딥러닝 기술들이 자연 언어 처리 분야에 적용되면서 문서 요약에 대해서도 다양한 방법론들이 제안되었다. 본 장에서는 최근의 자연어 생성 모델을 설명한다. 또 본 논문에서 제안하는 모델의 기반이 되는 모델을 소개한다.

2.1 시퀀스-투-시퀀스 주목 모델

기존의 시퀀스-투-시퀀스 모델(Sequence to Sequence Model)[2]의 문제 중 하나는 입력 값을 인코딩 하는 과정에서 전체 작업에서 필요가 없는 정보까지 인코딩 할 수밖에 없다는 문제가 존재한다. 입력 시퀀스의 길이가 길거나 포함되는 정보가 많은 경우, 그리고 선택적 인코딩이 불가능할 경우 문제가 발생한다.

시퀀스-투-시퀀스 주목 모델은 이러한 문제를 하기 위해 시퀀스-투-시퀀스 신경망에 주목 메커니즘을 결합한 모델이다. 주로 기계 번역, 문서 요약 분야에서 널리 쓰이고 있다. 시퀀스-투-시퀀스 주목 모델은 크게 다음과 같은 4가지 모듈(Module)으로 구성되어 있다.

1. 인코더 순환 신경망

인코더 순환 신경망(Encoder Recurrent Neural Network)은 요약의 목표가 되는 텍스트로부터 각 토큰을 읽고 이 토큰들의 임베딩을 입력으로 받는 모듈이다. 입력 시퀀스의 끝에 도달하면 순환 신경망의 마지막 은닉 상태를 디코더에 전달한다.

2. 디코더 순환 신경망

앞서 인코더로부터 전달받은 은닉 상태(Hidden State)를 이용해 디코더를 초기화한다. 요약글 구성하는 단어들의 시퀀스 형태로 결과값을 반환하게 된다. 디코더 순환 신경망(Decoder Recurrent Neural Network)은 인코더와 다르게 단방향 순환 신경망을 활용하며, 이전 단계의 단어들을 입력값으로 받게 된다.

3. 주목 분포와 문맥 벡터

대상 문서로부터 다음 단어를 생성할 때, 대상 문서의 단어에 대한 확률인 주목 분포는 인코더 순환 신경망과 디코더 순환 신경망의 은닉 상태를 입력 값으로 하는 함수로 표현된다. 이는 직관적으로 해당 신경망이 다음 순서의 단어를 생성 시 대상 문서의 어느 단어를 주목해서 봐야 할 지를 표현해 주는 것이다.

4. 어휘 분포

최종적으로 문맥 벡터와 디코더 은닉 상태(Hidden State)의 출력 값을 결합하여 2개의 선형결합 층을 통해서 단어의 분포를 표현한다. 이는 원문 텍스트가 아닌 전체 단어 요약을 통해 생성될 단어들을 순서대로 표현된다.

2.2 기억 신경망

기억 신경망(Memory Network)는 자연 언어 처리의 과제 중 텍스트 질의응답(Text Question Answering) 문제를 해결하고자 제안된 신경망 모델이다. 일반적으로 쓰이는 신경망 모델인 CNN, RNN은 분류나 번역 등의 과제에서는 좋은 성능을 보여준다. 그러나 입력 시퀀스 길이가 길어지는 질의 응답, 문서 요약 등의 문제에서는 취약한 성능을 보인다. 왜냐 하면 입력의 길이가 길어질수록 담아내야 할 정보의 크기가 커지므로 은닉 유닛의 크기가 커져야 하고, 데이터가 입력됨에 따라 입력 시퀀스 앞 부분의 위치한 정보들이 손실되기 때문이다. 이러한 문제를 해결하기 위해 제안된 모델이 기억 신경망[15] 모델이다. 기억 신경망은 기존의 신경망 이외에 별도의 외부 메모리를 갖고 있어 입력 시퀀스 길이에 상관 없이 오랜 시간 정보를 저장할 수 있고, 결과값을 생성할 때 학습 모델과 외부 메모리가 상호 작용할 수 있도록 설계했다.

본 연구에서는 기본적인 기억 신경망을 응용한 키-밸류 기억 신경망(Key-Value Memory Networks)[16]에서 영감을 얻었다. 키-밸류 기억 신경망의 기본 구조는 다음과 같다.

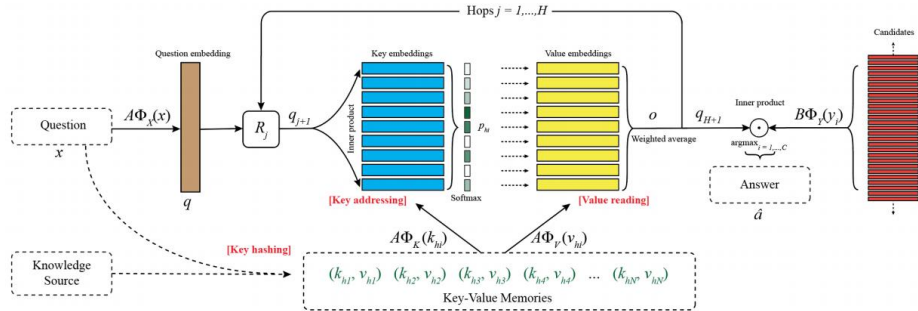


그림 1 키-밸류 기억 신경망의 구조

원래의 기억 신경망은 입력 메모리와 출력 메모리로 구성되어 있지만, 키-밸류 기억 신경망은 키 메모리와 밸류 메모리로 구성되어 있는 것이 특징이다.

제 3장 배경 지식

3장에서는 이 논문에서 기존 모델로 삼고 있는 문서 요약 신경망에 대해 설명하도록 한다.

3.1 문서 요약 신경망

먼저 다음 조건들을 만족하는 텍스트들의 쌍의 집합 (X, Y) 을 고려한다. X 는 대상 문서의 토큰들 x_1, \dots, x_n 대응되는 $x \in X$ 를 만족하는 문자열이고 Y 는 요약의 토큰들 y_1, \dots, y_m 에 대응되는 $y \in Y$ 를 만족하는 문자열이다. 여기서 n 과 m 은 $n \ll m$ 이다.

생성 요약은 요약을 생성할 때 한 번에 한 단어씩 생성한다. 이 때, 모델은 매 생성 단계마다 이전에 생성했던 단어들이 무엇인지를 알고 있다. 여기서 요약 신경망의 목표는 정확한 요약을 생성할 확률을 최대화하는 θ 를 모수(Parameter)로 가지는 $f(x)$ 를 학습하는 것이 문서 요약 신경망의 목표이다.

본 논문에서 기존 연구들을 토대로 시퀀스-투-시퀀스 주목 모델을 기존 모델로 설정했다. 신경망 내의 생성(Decoding) 단계 j 에서 주목 분포 $p(a_j | x, y_{1:j-1})$ 는 모든 대상 문서의 토큰들에 대한 느슨한 분포(Soft Distribution)를 표현하며, 이는 현재 모델이 어디를 주시하고 있는 지로도 해석할 수 있다.

또한 본 논문에서 사용한 기존 모델은 대상 문서에서 필요한 토큰을 가져오기 위해 복사 메커니즘[17]을 사용한다. 복사 메커니즘을 사용하는 경우 복사 확률 분포는 다음과 같은 결합 확률 분포로 표현할 수 있다.

$$p(y_j | y_{1:j-1}, x) = p(z_j = 1 | y_{1:j-1}, x) \times p(y_j | z_j = 1, y_{1:j-1}, x) + \\ p(z_j = 0 | y_{1:j-1}, x) \times p(y_j | z_j = 0, y_{1:j-1}, x)$$

이 결합 확률 분포는 z_j 의 값에 따라 각 단계별로 요약을 위한 단어를 생성할 지 아니면 대상 문서에서 복사를 할 지에 대한 여부가 결정된다.

제 4장 추가 정보를 고려한 문서 요약 모델

이번 장에서는 요약의 목표가 되는 문서 외에 별도의 추가 정보를 이용해 문서 요약의 질을 향상시키는 문서 요약 모델을 제안한다. 1절에서는 기존 문서 요약 연구들에 대한 한계점을 살펴보고, 2절에서는 이를 해결하기 위한 추가 정보를 위한 추가 정보 인코딩 방법론을 제안한다. 그리고 마지막 3절에서는 지식 인코딩을 기존 모델에 어떻게 결합시키는 방법을 제안한다.

4.1 기존 연구의 한계점

앞서 소개했던 문서 요약 모델들은 모두 CNN/Daily Mail 데이터셋을 학습 데이터를 기준으로 성능을 평가하고 있다. 이 CNN/Daily Mail 데이터 셋의 경우 CNN과 Daily mail의 뉴스 데이터들을 모아 놓은 데이터 셋으로, 총 287,226의 학습 셋과 13,368의 검증 셋, 11,490의 평가 셋으로 구성되어 있다. 또한 뉴스를 기반으로 한 데이터 셋이기 때문에 문서 텍스트의 품질이 매우 우수하다. 또한 뉴스의 목적을 고려하면 뉴스 자체가 완결된 문서이므로 요약을 하는데 있어 별도의 추가 정보가 필요하지 않는 경우가 많다. 그러나 현실의 데이터들은 이렇게 양이 많지도, 질이 좋지 않다. 기존의 문서 요약들은 현실의 텍스트 데이터에 대해 잘 다루려고 하지 않으려고 한다. 그렇기 때문에 현실의 데이터들에 대해서 충분히 검증되지 않았음을 알 수 있다.

그러므로 본 연구에서는 실제 데이터들에 대해 대표적인 기존 모델의 요약 성능을 검증하고 추가 정보를 이용해 기존 모델보다 어떻게 더 나은 성능을 낼 수 있는지에 대해 제안하도록 한다.

4.2 추가 정보 인코더

먼저 대상 문서와 같이 입력으로 들어오는 추가 정보의 형식에 대해 정의한다. 입력으로 들어오는 추가 정보는 고정된 항목들과 그 항목에 대한 텍스트들이 나열되어 있다고 가정한다. 또한 추가 정보의 항목 수는 모든 대상 문서에 대해 동일한 크기를 가지는 것으로 가정한다. 각 항목의 추가 정보는 키(Key)와 밸류(Value)의 쌍인 (K, V) 로 정의한다. K 는 추가 정보의 항목이 가지는 개별적인 키고, $\{k_1, \dots, k_n\} \in K$ 를 만족하는 개별 키 k 들을 정의할 수 있다. 여기서 n 은 추가 정보 항목의 숫자이다. V 는 각 항목이 가지고 있는 추가 정보 토큰들의 집합으로, 각 키에 들어가 있는 값들을 표현한다. 또한 키 k_i 에 포함된 추가 정보 토큰들은 $\{v_{i1}, \dots, v_{im}\}$ 으로 표현할 수 있다.

앞서 소개했던 키-밸류 기억 신경망과는 다르게 추가 정보 항목의 크기가 고정되어 있으므로 별도의 키 임베딩을 사용하지 않는다. 추가 정보들의 토큰들 $\{v_{i1}, \dots, v_{im}\}$ 을 각각 BoW(Bag-of-Words) 기법을 이용해 One-hot vector $\{x_{i1}, \dots, x_{im}\}$ 로 만든다. 그리고 이 벡터들을 밸류 임베딩 행렬(Value embedding matrix)를 통해 인코딩한다. 여기서 요약 문서의 입력 토큰과 추가 정보들의 토큰은 같은 단어라고 하더라도 시멘틱(Semantic)이 다르다고 생각해서 임베딩 공간을 공유하지 않는 것으로 설계했다. 또한 별도의 미리 학습된 임베딩을 사용하지 않고 랜덤하게 초기화된 임베딩 행렬을 사용했다.

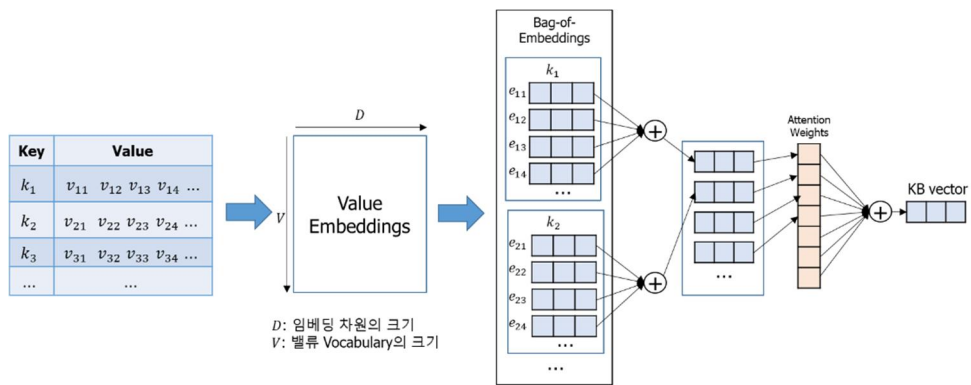


그림 2 추가 정보 인코더의 구조

이러한 과정을 식으로 표현하면 다음과 같다.

$$k'_i = \sum A x_{ij}$$

여기서 k'_i 는 키 k_i 의 메모리 벡터(Memory vector)를 의미하고, A 는 벡터를 임베딩하기 위한 행렬이다. 이를 통해 각각의 키 k_i 에 대한 메모리 벡터들을 생성할 수 있다. 이 메모리 벡터는 각 키들에 대응되는 추가 정보의 임베딩으로 볼 수 있다.

4.3 기존 모델과 결합

현재 널리 쓰이는 먼저 소개했던 요약 신경망 모델의 구조는 다음과 같다.

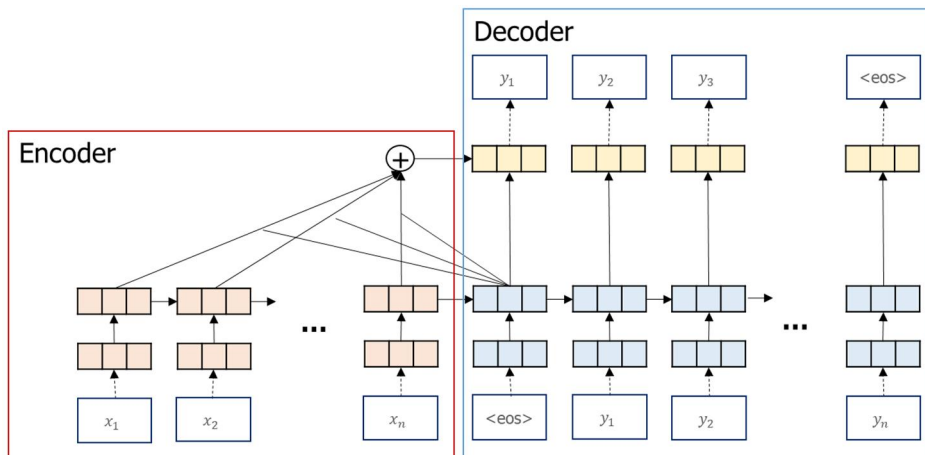


그림 3 요약 신경망의 기본적인 구조

이 모델은 기본적인 시퀀스-투-시퀀스 모델에 주목 메커니즘이 포함된 모델이다. 앞서 제안했던 추가 정보 임베딩을 위 구조에 어떻게 적용할 지 본 절에서 살펴보도록 한다.

1. 추가 정보 별 가중치 설정

각 대상 문서에 대해 필요한 추가 정보가 다를 것이다. 그렇기 때문에 대상 문서가 어떠한 추가 정보가 필요한지에 따라 추가 정보 각 항목별로 얼마만큼의 비중을 주어야 하는지 판단해야 한다. 본 모델에서는 이러한 비중을 구하기 위해 다음과 같은 구조를 제시한다.

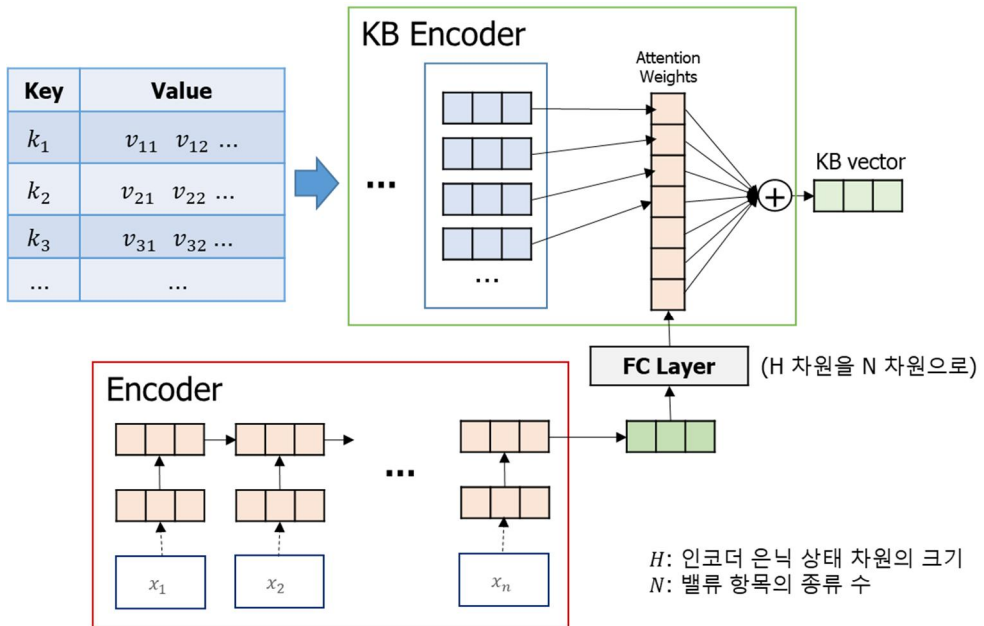


그림 4 추가 정보 임베딩 생성 과정

입력 시퀀스 끝나면, 인코더의 마지막 은닉 상태가 디코더에 전달된다. 이 은닉 상태는 요약 문서가 인코딩을 거쳐간 후 생성된 결과물로서 입력 시퀀스의 정보가 내재되어 있는 벡터로 볼 수 있다. 일반적인 시퀀스-투-시퀀스 모델에서는 이 벡터를 디코더에 전달해서 요약을 생성한다. 본 논문에서 제안하는 모델에서도 디코더에 전달하는 부분은 동일하지만, 입력 시퀀스의 정보가 내재되어 있다는 점을 이용해 추가 정보 선별에 사용한다.

먼저 벡터를 [은닉 상태 차원 크기] * [밸류 항목의 종류 수] 크기의 완전연결 레이어(fully Connected Layer)를 통과시킨다. 여기서 완전 연결 레이어도 학습에 포함시켜서, 대상 문서에 따라 각 항목마다 얼마만큼 비중을 부여할 지 결정한다. 이렇게 통과한 벡터에 대해 소프트맥스(Softmax) 함수를 통과시켜 분포화 시켰다. 그리고 마지막으로 이렇게 분포화된 벡터를 가중치로 삼아 미리 인코딩 된 추가 정보 항목 벡터들을 가중치 합을 통해 하나로 합쳤다. 이를 통해 입력 시퀀스가 각각의 추가 정보 항목의 필요한 정도를 결정할 수 있게 하도록 의도했다.

2. 추가 정보 임베딩 결합

마지막으로 인코딩된 추가 정보 임베딩을 다음과 같이 디코더에 결합해준다.

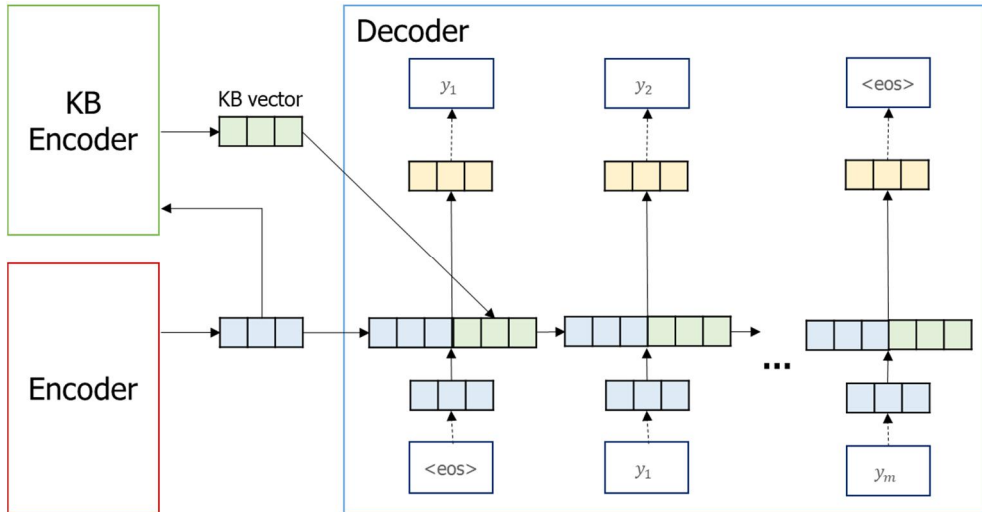


그림 5 추가 정보 임베딩이 추가된 디코더 구조

디코더의 은닉 유닛에 추가 정보 임베딩을 접합(Concatnation)시킨다. 그리고 주목 벡터의 차원도 같이 증가시켜서 단어 생성 확률 분포에 추가 정보가 반영될 수 있도록 모델을 수정한다. 이러한 구조를 통해 추가 정보가 요약 생성에 영향을 주는 것을 기대할 수 있도록 의도했다.

제 5장 실험 방법 및 결과

4장에서는 본 논문에서 제안한 추가 정보를 이용한 문서 요약 모델의 성능을 평가한다. 1절에서는 데이터셋과 전처리에 대해 설명한다. 2절에서는 실험 설계에 대해 설명하고 3절에서는 실험 결과에 대한 분석을 진행하고 4절에서는 추가 정보가 요약에 어떤 영향을 주는지 그 결과에 대해 살펴본다.

5.1 데이터 셋

5.1.1 학습 데이터 셋

요약이 필요한 대상 문서와 추가 정보가 같이 존재하는 데이터는 여러 가지 종류가 있지만, 그 중 상품 후기 데이터를 실험에 사용하기로 결정했다. 상품 후기를 대상 문서로 설정하고 상품 후기의 제목을 요약으로 가정했다.

본 연구에서 문서 요약에 사용하기 위해 패션 쇼핑몰 무신사(Musinsa)¹에서 상품 리뷰를 1,265,311건 수집했다. 그 중 대상 문서와 요약 길이를 조건으로 161,759건의 항목을 추출했다. 데이터의 상세는 <표 1>와 같다.

표 1 수집한 데이터 셋에 대한 정보

| 총 항목수 | 대상 문서의 평균 단어 수 | 요약의 평균 단어 수 |
|----------|----------------|-------------|
| 161,759개 | 3.27개 | 17.57개 |

¹ <https://store.musinsa.com>

대상 문서와 요약 이외에 모델에 사용하기 위해 수집한 추가 정보는 <표 2>와 같다.

표 2 상품 후기 데이터 세트의 추가 정보 표

| 항목 | 설명 | 누락값 수 | 누락값 비율 |
|-----------|-----------|--------|---------|
| Brand | 상품의 브랜드 명 | 9,238 | 5.71 % |
| Category | 상품의 카테고리 | 0 | 0 % |
| Title | 상품의 제목 | 0 | 0 % |
| Sub title | 상품의 부제 | 3 | 0.001 % |
| Season | 상품의 시즌 | 73,373 | 45.36 % |
| Gender | 상품의 대상 성별 | 73,373 | 45.36 % |
| Rate | 상품 후기의 평점 | 0 | 0 % |

위와 같은 형식으로 수집한 데이터의 한 항목의 예시는 <표 3>와 같다.

표 3 상품 후기 데이터 항목 예시

| 항목 | 예시 |
|-----------|--|
| 대상 문서 | 오버핏이고 따뜻하고 재질도 좋아요 조기배송에 레이어드 입으면 이뻐요 넘 편해서 운동용으로 검은색도 추가구매할려구용 |
| 요약 | 오버핏이고 따뜻하고 재질도 좋아요 |
| Brand | HALBKREIS |
| Category | 맨투맨 스웨트셔츠 상의 |
| Title | 기모 지퍼맨투맨 TA-17_GA |
| Sub title | TA-17 Zipper Sweat Shirt_GA |
| Season | 2017 F/W |
| Gender | 남 |

| | |
|------|---|
| Rate | 5 |
|------|---|

5.1.2 데이터 전처리

본 연구에서는 수집한 데이터들은 가공되지 않은 텍스트들이라서 부정확한 띄어쓰기, 이모티콘 등의 특징이 있다. 이러한 특징들이 문서 요약 학습을 방해하는 노이즈가 될 수 있기 때문에 이모티콘을 비롯한 특수 문자와 ‘ㄴ’ 와 ‘ㄷ’ 등의 문자를 제거하는 데이터 전처리(Data Preprocessing) 과정을 진행했다. 또한 띄어쓰기 문제 해결 및 토큰화를 위해 Komoran² 라이브러리를 이용해서 형태소 단위로 파싱(Parsing)을 진행했다. 전처리 전후의 결과물은 다음 <표 4>와 같다.

표 4 데이터 전처리 예시

| 항목 | 예시 |
|--------|---|
| 원문 | 바지 되게 편하고 따뜻해요 ㄴ ㄴ 독서실갈때나 도서관갈때 자주입는데 요즘같은 날씨에 입기 좋아요! |
| 전처리 결과 | 바지 되 게 편 하 고 따듯 하 어요 독서실 갈 때 나 도서관 갈 때 자주 입 는데 요즘 같 은 날씨 에 입 기 좋 아요 |

실험에 사용한 데이터 세트는 155,759개의 학습 세트와 5,000개의 검증 세트, 1,000개의 평가 세트로 구성하였다.

5.2 실험 설계

앞서 구축한 데이터 세트에 대해 다음과 같이 네가지 모델에 대해 실험을 진행했고 그 결과를 여러 가지 관점으로

² <https://www.shineware.co.kr/products/komoran/>

비교해보았다.

1. 기존 모델(Baseline Model)

일반적으로 문서 요약에 사용되는 시퀀스-투-시퀀스 모델에 주목 메커니즘을 적용한 모델을 기존 모델로 정했다.

2. 기존 모델에 추가 정보 접합

그 다음으로는 입력 시퀀스의 뒤에 추가 정보를 추가시켜서 기존 모델에 학습시키고 그 결과를 확인했다. 이는 단순히 기존 모델에 추가 정보를 추가시키는 것과 추가 정보를 인코딩해서 추가하는 것의 차이를 확인하기 위해서 설정한 모델이다.

3. 추가 정보 인코딩 모델

본 논문에서 제시한 추가 정보를 별도의 인코딩을 통해 추가 정보 임베딩을 만들고 그 임베딩을 디코더에 추가시키는 모델을 사용했다. 추가 정보의 개수에 따라 각각 실험을 진행했는데 첫 번째로는 데이터 셋의 수집한 추가 정보 7개 모델을 사용하는 모델을 실험 모델로 선정했다. 그리고 누락 값이 많은 2개(시진, 성별)의 추가 정보 항목이 오히려 학습과 생성에 노이즈로 작용할 수 있다고 생각해서 해당 항목을 제외한 추가 정보 5개만을 사용하는 모델 또한 실험 모델에 포함시켰다.

실험에 사용한 기기의 GPU는 NVIDIA GeForce GTX 1080을 사용했으며, 구현 언어는 Python, 딥 러닝(Deep Learning) 라이브러리(Library)로는 Pytorch³를 사용했다. 그리고 대상 문서와 요약의 어휘들의 단어 임베딩은 미리 학습된 FastText

³ <https://pytorch.org/>

한글 임베딩⁴ 값을 사용했다.

5.3 실험 결과 및 분석

학습한 모델을 통해 생성한 요약과 평가 세트의 요약에 대상으로 블루 스코어(BLEU Score)를 통한 평가와 사용자 평가(Human Evaluation)을 진행하였다.

I. 블루 스코어

문서 요약 성능의 자동 평가에는 여러 가지 방법이 있는데, 그중 많이 사용하는 방법 중 하나는 블루(BLEU: Bilingual Evaluation Understudy) 스코어이다. 이는 모델이 생성한 문장과 정답 문장 간의 정확도(Precision)를 평가하는 척도이다. 블루 스코어를 구하는 공식은 다음과 같다.

$$BLEU = \min \left(1, \frac{\text{output length}(\text{예측 문장})}{\text{reference length}(\text{실제 문장})} \right) \left(\prod_{i=1}^4 \text{precision}_i \right)^{\frac{1}{4}}$$

각 모델이 생성한 요약과 평가 세트의 요약의 블루 스코어 값은 다음의 <표5>와 같다.

⁴ <https://fasttext.cc/docs/en/crawl-vectors.html>

표 5 각 모델의 블루 스코어

| Model | BLEU |
|-------------|-------|
| 기준 모델 | 9.64 |
| 추가 정보 접합 | 10.51 |
| 추가 정보(7 항목) | 11.86 |
| 추가 정보(5 항목) | 11.51 |

추가 정보를 포함시켜서 같이 학습시킨 모델이 기준 모델보다 더 좋은 블루 스코어를 내는 것을 알 수 있었다.

2. 사용자 평가(Human Evaluation)

블루 스코어는 생성 모델이 생성한 결과와 정답이 얼마나 많이 겹치는지를 가지고 평가하는 척도이다. 즉, 요약 모델에서는 생성한 요약과 평가 세트의 요약의 토큰이 몇 개나 겹치는지를 가지고 평가를 한다. 이는 생성한 요약과 평가 세트의 요약이 같은 의미(Semantic)를 가지지만 다른 방식으로 표현될 때는 정확하지 않은 평가가 된다는 문제가 있다. 이를 극복하기 위해 각각의 생성된 요약에 대해 사용자 평가를 진행하였다.

사용자 평가는 각 요약에 대해 2가지 척도로 채점을 한다. 첫번째 척도는 연관성(Relevance)로 생성한 요약이 얼마만큼 본문의 정보를 잘 담고 있는가를 평가하는 척도이다. 두번째 척도는 가독성(Readability)로 생성한 요약이 얼마만큼 잘 생성되었는지를 평가하는 척도이다. 평가 세대에서 각 모델에서 생성한 요약 100개를 대상으로 총 5명에게서 각 요약에 대해 평가를 실시하였다. 사용자 평가의 결과는 다음 <표 6>와 같다.

표 6 사용자 평가 결과

| 모델 | 관련성 | 가독성 |
|-----------|------|------|
| 기준 모델 | 3.53 | 4.65 |
| 추가 정보 접합 | 3.25 | 4.49 |
| 추가 정보(7개) | 3.62 | 4.55 |
| 추가 정보(5개) | 3.67 | 4.79 |

추가 정보를 5개만 사용한 것이 관련성과 가독성에서 가장 좋은 결과를 내었음을 확인할 수 있었다. <표 7>이 각 모델과 추가 정보에 따른 요약의 결과이다.

표 7 각 모델별 요약 결과

| 항목 | 내용 |
|-------|---|
| 대상 문서 | 피로도 너무너무 좋고 워단도 좋고 다 좋으데 배송이 한달걸림 시켜놓고 까먹었는데 와서 반가웠네요 |
| 추가 정보 | vastic /셔츠 블라우스 상의 / 코튼 워크셔츠 베이지 / vastic cotton work shirts beige / all all / 남 / 5 |

| | | |
|----------|----|---------------|
| 요약(제목) | | 진 짜너머너무너무 이쁜데 |
| 기준 모델 | | 다 좋은데 배송이 |
| 추가 정보 접합 | | 다 좋은데 배송이 |
| 추가 정보 | 5개 | 배송 빠르고 다 좋아요 |
| 인코딩 | 7개 | 너무 너무 좋아요 |

5.4 추가 정보 영향 평가

5.4.1 추가 정보 가중치 실험

1,000개의 평가 셋에 대해, 각각의 추가 정보가 가지는 가중치들의 평균을 계산해보았다. 이는 추가 정보들이 각각 얼마의 비율만큼 반영되는지 알기 위해서이다. 그 결과는 아래 <표 8>과 같다.

표 8 추가 정보 항목별 가중치 평균

| 항목 | 가중치 평균 |
|-----------|--------|
| Brand | 6.9% |
| Category | 15.7% |
| Title | 11.4% |
| Sub title | 19.3% |
| Season | 22.8% |
| Gender | 7.4% |
| Rate | 16.1% |

예상한 것과 다르게 시진과 부제목이 가장 큰 비중을 차지하고 있었다. 그 다음으로 영향을 많이 주는 추가 정보 항목은 사용자 평점으로 이는 대상 문서의 극성을 요약에 더 잘 반영하기 위해 많은 가중치를 가져가도록 학습했음을 알 수 있었다.

5.4.2 추가 정보 제어 실험

이번 절에서는 추가 정보가 실제로 요약 생성에 영향을 주는지 확인해보는 실험을 진행하였다. 먼저 사용자 평점(Rate) 정보의 특성은 구매자가 상품에 대한 전체적인 감성을 대표하는 정보이다. 1, 2점의 경우 구매자가 상품에 대해 부정적인(Negative) 감정 극성을 가진다고 볼 수 있고, 반대로 4, 5점의 경우 구매자가 상품에 대해 긍정적인(Positive) 감정 극성을 가진다고 볼 수 있다.

이번 실험에서 살펴보고자 하는 3점의 경우, 구매자는 상품에 대해 복합적인(Complex) 감정 극성을 가지고 있으며, 대상 문서에 상품의 장점과 단점이 모두 포함되어 있을 것이라 예상된다.

이를 이용해서 추가 정보가 요약에 주는 영향을 확인하기 위해, 사용자 평점이 3인 요약 항목들을 추출한 후, 사용자 평점 값을 직접 1~5점으로 변화시켜가면서 모델에 입력했을 때 생성된 요약을 확인했다. 그 결과 중 하나는 아래 <표 9>와 같다.

| 대상 문서 | | 두께도 사이즈도 좋습니다 179 70 L 잘맞아요 근데 문제가 냄새가 너무 나네요 배판다에 걸어놓고 냄새 났는데 한참 걸리네요 |
|--------|---|---|
| 사용자 평점 | 1 | 냄새가 너무 심하네요 |
| | 2 | 냄새가 너무 나네요 |
| | 3 | 냄새가 너무 나네요 |
| | 4 | 옷은 괜찮은데 냄새가 |
| | 5 | 사이즈도 좋고 이쁘네요 |

표 9 추가 정보 제어 요약 생성 결과

먼저 후기를 살펴보면 ‘두께도 사이즈도 좋습니다’ 의 긍정적인 내용과 ‘냄새가 너무 나네요’ 라는 부정적인 의견이 혼합되어 있는 것을 확인할 수 있다. 그리고 실험 결과를 살펴 보면, 별점에 따라 1,2,3점은 부정적인 내용을 위주로 요약을 했고, 5점의 경우 긍정적인 내용만을 요약했음을 확인할 수 있다. 이를 통해 간접적으로나마 추가 정보가 요약 을 생성하는데 영향을 주었음을 확인할 수 있다.

5.4.3 추가 정보 영향력 실험

이번 절에서는 각각의 추가 정보가 어떤 식으로 리뷰 요약에 도움을 주는지 실험을 해보다. 먼저 추가 정보를 요약에 영향을 주는 다음과 같이 2개의 그룹으로 나눴다.

표 9 비교를 위한 추가 정보 분할

| 그룹 1 | 그룹 2 |
|---|--------------------------|
| Brand Category Title Sub title | Season Gender Rate |

그룹 1에는 상품 요약에 직접적으로 도움이 될 거라 생각하는 추가 정보들을 분류했고, 그룹 2에는 그룹 1에 비해 상대적으로 도움이 되기 어려운 추가 정보들로 묶었다. 각 그룹별 요약 결과 예시는 아래 <표 10>과 같다.

표 10 각 그룹별 요약 결과 예시

| 항목 | 내용 |
|-------|---|
| 대상 문서 | 제 생각보다 두꺼지지 않아서 봄에 이거만 입고 다니기 좋을 거 같아요 새끼들도 뭔가 새파라해서 더 마음에 드네요 |

| | | |
|--------|------|---|
| 추가 정보 | | costume o' clock / 맨투맨 스웨트셔츠 상의 / 티25에이치 에프 스웨트 셔츠 블루 / t25h f sweat shirt orange / 2018 spring season / 남 / 5 |
| 요약(제목) | | 새가일이 이쁘네요 |
| 추가 정보 | 그룹 1 | 새가일이 이쁘네요 |
| 인코딩 | 그룹 2 | 봄에 입기 좋은 거 같아요 |

위의 <표 10>을 살펴보면, 그룹 2의 계절 정보가 요약 결과에 반영되었음을 확인할 수 있었다.

제 6장 결론 및 향후 연구

6.1 결론

본 논문에서는 기억 신경망을 응용해 추가 정보를 학습에 포함시키는 문서 요약 모델을 제시하였다. 문서 요약에 대한 기존의 많은 연구들이 정제된 데이터를 기반으로 연구가 진행되어 왔지만, 이는 현실의 데이터들에 적용하는데 있어서 한계가 있다는 문제점을 지닌다.

이러한 문제들을 해결하기 위해 추가 정보를 인코딩하는 방법을 제안했다. 이렇게 인코딩한 임베딩을 기존 모델에 적용시키는 기법을 제시했다. 이를 통해 쇼핑물 리뷰 데이터셋에서 기존 모델보다 요약 대상인 상품 리뷰와 연관된 추가 정보를 통해 요약의 품질이 향상된 것을 확인할 수 있었다.

6.2 향후 연구

향후에는 다음과 같은 사항들에 대해 연구 및 보완할 계획이다. 첫째로는 본 연구에서는 추가 정보를 임베딩으로 인코딩하고, 인코딩된 임베딩들의 가중치 합으로 추가 정보 임베딩을 생성하였다. 하지만 이런 방법 외에도 추가 정보를 인코딩하는 많은 방법들이 있다. 추후에는 이러한 방법들을 적용해서 본 논문의 모델과 성능을 비교해볼 수 있을 것이다.

또한 본 연구는 디코딩 단계에서 대상 문서의 정보와 추가 정보의 비중을 동일하게 가져가지만, 별도의 선별(Gating)을 통해 원문과 추가 정보의 비중을 조절해서 실험해볼 수 있을 것이다. 현재는 모든 학습과 생성을 모델에 위임한 상태지만 선별의 패러미터를 조절함으로써 더 좋은 품질의 요약물 생성해낼 수 있을 것이라 생각한다.

본 연구에서는 추가 정보를 인코딩한 임베딩을 디코더의 초기화에서만 사용하지만 이는 요약의 길이가 길어질수록 정보가 손실되는 문제가 있다. 이러한 문제를 해결하기 위해 각 디코딩 단계별로 추가 정보 인코딩을 적용하는 구조를 생각해볼 수 있다. 이를 통해 요약 길이가 길어지더라도 추가 정보가 더 잘 반영되므로 더 좋은 성능을 기대할 수 있을 것이다.

평가 척도에 대한 추가 연구도 생각해볼 수 있다. 문서 요약 모델의 성능 평가에 일반적으로 BLEU 스코어나 ROUGE 스코어 등의 평가 척도를 사용하지만, 이러한 평가 척도들은 대상 문서에서 어떤 내용을 선택했는지만을 평가하고 일관성, 유창성, 문법 등과 같은 다른 품질 측면을 고려하지 않는다는 문제가 있다. 그래서 최신의 문서 요약 연구들은 상이한 평가 척도 외에도 사용자 평가로 별도의 성능 평가를 하는 것을 확인할 수 있다. 이런 문제를 해결하기 위해 자동 평가 척도에 대한 추가적인 연구를 통해 문서 요약 분야에 도움을 줄 수 있을 것이라 생각한다.

또한 ‘문서 요약’에 대한 평가 척도와 ‘리뷰 요약’에 대한 평가 척도는 다른 평가 기준을 통해서 정해져야 할 것이다. 본 연구에서는 충분히 다루지 못했지만 요약의 ‘목적성’이 충분히 반영된 평가 척도에 대한 연구도 큰 도움이 될 수 있을 것이라 생각한다.

참고 문헌

- [1] Chen, Danqi, Jason Bolton, and Christopher D. Manning. "A thorough examination of the cnn/daily mail reading comprehension task." arXiv preprint arXiv:1606.02858 (2016).
- [2] Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. "Sequence to sequence learning with neural networks." *Advances in neural information processing systems*. 2014.
- [3] Cho, Kyunghyun, et al. "Learning phrase representations using RNN encoder-decoder for statistical machine translation." arXiv preprint arXiv:1406.1078 (2014).
- [4] Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." arXiv preprint arXiv:1409.0473 (2014).
- [5] Weston, Jason, Sumit Chopra, and Antoine Bordes. "Memory networks." arXiv preprint arXiv:1410.3916 (2014).
- [6] Miller, Alexander, et al. "Key-value memory networks for directly reading documents." arXiv preprint arXiv:1606.03126 (2016).
- [7] Sukhbaatar, Sainbayar, Jason Weston, and Rob Fergus. "End-to-end memory networks." *Advances in neural information processing systems*. 2015.
- [8] Paice, Chris D. "Constructing literature abstracts by computer: techniques and prospects." *Information Processing & Management* 26.1 (1990): 171-186.
- [9] Kupiec, Julian, Jan Pedersen, and Francine Chen. "A trainable document summarizer." *Advances in Automatic Summarization* (1999): 55-60.
- [10] Erkan, Gjjnes, and Dragomir R. Radev. "Lexrank: Graph-based lexical centrality as salience in text summarization." *Journal of artificial intelligence research* 22 (2004): 457-479.
- [11] Genest, Pierre-Etienne, and Guy Lapalme. "Fully abstractive approach to guided summarization." *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 2012.
- [12] See, Abigail, Peter J. Liu, and Christopher D. Manning. "Get to the point: Summarization with pointer-generator networks." arXiv preprint arXiv:1704.04368 (2017).
- [13] Gehrmann, Sebastian, Yuntian Deng, and Alexander M. Rush. "Bottom-up abstractive summarization." arXiv preprint arXiv:1808.10792 (2018).

- [14] Fattah, Mohamed Abdel, and Fuji Ren. "GA, MA, FFNN, PNN and GMM based models for automatic text summarization." *Computer Speech & Language* 23.1 (2009): 126-144.
- [15] Weston, Jason, Sumit Chopra, and Antoine Bordes. "Memory networks." *arXiv preprint arXiv:1410.3916* (2014).
- [16] Miller, Alexander, et al. "Key-value memory networks for directly reading documents." *arXiv preprint arXiv:1606.03126* (2016).
- [17] Vinyals, Oriol, Meire Fortunato, and Navdeep Jaitly. "Pointer networks." *Advances in Neural Information Processing Systems*. 2015.

Abstract

A Product Review Summarization Considering Additional Information

Jaeyeun Yoon

Department of Computer Science and Engineering

The Graduate School

Seoul National University

Automatic document summarization is a task of creating a document in a form that is suitable for a particular user or task from a given document. With the development of the Internet, various kinds of data including texts are explosively increasing, and in this situation, the value of document summarization is increasing. It also has important value as a way of studying how machines understand language.

In a previous study, document summarization using graph-based model or statistical based model were common. However document summarization models using various deep-learning models are mainly studied.

Although these various models show good summarization performance comparable to a human-generated summary, there is a problem that the performance depends on the quantity and quality of the training data.

Text to be summarized in the document generally includes news articles, websites, and product reviews. Most studies are based on news articles that are both good in terms of both quantity and quality of training data. In this study, we will focus on summary techniques for product reviews that have relatively poor quality and quantity of training data.

Therefore, in this study, we propose a document summarization model that provides a better quality summary by adding additional information about given product reviews to solve the problem of the quantity and quality of such data. In this paper, we compare the results of the proposed model with the summaries generated by baseline model, and the results of various experiments including human evaluation, the quality of the product review summary proposed in this study showed better quality of both relevance and readability than the other models, Experiments have shown how additional information can affect document summaries.

Keywords: Document summarization, Sequence-to-sequence model, Attention mechanism, Memory network, Artificial neural network

Student Number: 2016-26565